1. **MACHINE LEARNING**
2. Định nghĩa về Machine Learning

Machine Learning (Học máy) là thuật ngữ chỉ các hành động dạy máy tính để cải thiện một nhiệm vụ mà nó đang thực hiện. Cụ thể khả năng của Machine Learning là sử dụng các thuật toán để phân tích những thông tin sẵn có, sau đó học hỏi và đưa ra những quyết định, dự đoán về những thứ có liên quan. Thay vì phải lập trình ra một phần mềm với các hành động, hướng dẫn chi tiết để thực hiện một nhiệm vụ thì máy tính sẽ được học cách thực hiện các nhiệm vụ đó thông qua một lượng dữ liệu và các thuật toán.

Machine learning đang trở thành một yếu tố quan trọng trong lĩnh vực khoa học dữ liệu đang phát triển. Bằng cách sử dụng các phương pháp thống kê và đào tạo các thuật toán, machine learning có khả năng cung cấp các phân loại hoặc dự đoán và khám phá thông tin chi tiết từ những dự án khai thác dữ liệu.

Sử dụng thông tin chuyên sâu để thúc đẩy các quyết định về ứng dụng và kinh doanh cũng như tác động đến các chỉ số tăng trưởng. Khi dữ liệu lớn mở rộng và phát triển, nhu cầu về các nhà khoa học dữ liệu cũng tăng theo. Họ được yêu cầu giúp xác định các câu hỏi kinh doanh chính và thông tin cần thiết để trả lời chúng.

Thường thì các bài toán trong machine learning được phân thành hai loại chính: dự đoán (prediction) và phân loại (classification). Các bài toán dự đoán thường liên quan đến việc ước lượng giá trị, như giá nhà, giá xe, trong khi các bài toán phân loại thường liên quan đến việc xác định danh tính hoặc loại của một đối tượng, như nhận diện chữ viết tay, đồ vật, và các nhiệm vụ tương tự.

1. Các phương pháp Machine Learning

* **Supervised machine learning (Học Có Giám Sát)**

Học máy có giám sát, hay còn gọi là máy học có giám sát, là phương pháp sử dụng các tập dữ liệu được gắn nhãn để huấn luyện các thuật toán phân loại dữ liệu hoặc dự đoán kết quả một cách chính xác. Mô hình được điều chỉnh trọng lượng của nó khi dữ liệu đầu vào được đưa vào, trong quá trình xác nhận chéo để tránh tình trạng quá mức hoặc thiếu thông tin.

Học máy có giám sát hỗ trợ tổ chức giải quyết nhiều vấn đề thực tế quy mô lớn, như phân loại thư rác trong hộp thư đến cá nhân.

Các phương pháp thường được sử dụng trong Học máy có giám sát bao gồm: logistic regression, neural networks, linear regression, naive bayes, random forest và support vector machine (SVM).

* **Unsupervised machine learning (Học Không Giám Sát)**

Học máy không giám sát là phương pháp sử dụng các thuật toán máy học để phân tích và phân cụm các tập dữ liệu không được gắn nhãn. Không yêu cầu sự can thiệp của con người, các thuật toán này có khả năng phát hiện mẫu hoặc nhóm dữ liệu ẩn, làm cho chúng lý tưởng cho việc phân tích dữ liệu khám phá và nhận diện hình ảnh và mẫu.

Học máy không giám sát cũng được sử dụng để giảm số lượng tính năng trong một mô hình thông qua quá trình giảm kích thước, với phương pháp như phân tích thành phần chính và phân tích giá trị đơn lẻ.

Các thuật toán thường được sử dụng trong Học máy không giám sát bao gồm: k-means clustering, neural networks và các phương pháp phân cụm xác suất.

* **Semi-supervised learning (Học Tăng Cường)**

Học máy bán giám sát cung cấp một phương pháp hiệu quả giữa học máy có giám sát và không giám sát. Trong quá trình đào tạo, nó sử dụng một tập dữ liệu có nhãn nhỏ hơn để hướng dẫn phân loại và trích xuất tính năng từ một tập dữ liệu lớn hơn, không được gắn nhãn. Phương pháp này có thể giải quyết vấn đề thiếu dữ liệu được gắn nhãn cho thuật toán học có giám sát và làm giảm chi phí đánh nhãn đối với dữ liệu.

* Deep Learning (Học Sâu)

Học sâu là một phần của học máy, đặc biệt là mạng nơ-ron sâu có ba lớp trở lên. Những mạng nơ-ron này cố gắng mô phỏng hành vi của não người, giúp deep learning "học" từ lượng lớn dữ liệu. Công nghệ này đã thúc đẩy nhiều ứng dụng và dịch vụ trí tuệ nhân tạo như trợ lý kỹ thuật số, điều khiển từ xa hỗ trợ giọng nói và phát hiện gian lận thẻ tín dụng.

Công nghệ deep learning được áp dụng rộng rãi trong các sản phẩm và dịch vụ hàng ngày như trợ lý kỹ thuật số, điều khiển từ xa hỗ trợ giọng nói và phát hiện gian lận thẻ tín dụng, cũng như trong các lĩnh vực mới nổi như ô tô tự lái.

1. Quy trình hoạt động của máy học

Quy trình làm việc với machine learning gồm 5 bước cụ thể từng bước trong machine learning như sau:

Data collection – thu thập dữ liệu: để máy tính có thể học được bạn cần có một bộ dữ liệu (dataset), bạn có thể tự thu thập chúng hoặc lấy các bộ dữ liệu đã được công bố trước đó. Lưu ý là bạn phải thu thập từ nguồn chính thống, có như vậy dữ liệu mới chính xác và máy có thể học một cách đúng đắng và đạt hiệu quả cao hơn.

Preparing data **-** chuẩn bị dữ liệu*:* bước này dùng để chuẩn hóa dữ liệu, loại bỏ các thuộc tính không cần thiết, gán nhãn dữ liệu, mã hóa một số đặc trưng, trích xuất đặc trưng, rút gọn dữ liệu nhưng vẫn đảm bảo kết quả… Bước này tốn thời gian nhất tỉ lệ thuận với số lượng dữ liệu bạn có. Bước 1 và 2 thường chiếm hơn 70% tổng thời gian thực hiện.

Training model – huấn luyện mô hình: bước này là bước huấn luyện cho mô hình hay chính là cho nó học trên dữ liệu đã thu thập và xử lý ở hai bước đầu.

Evaluating model – đánh giá mô hình: sau khi đã huấn luyện mô hình xong, cần dùng các độ đo để đánh giá mô hình, tùy vào từng độ đo khác nhau mà mô hình cũng được đánh giá tốt hay không khác nhau. Độ chính xác của mô hình đạt trên 80% được cho là tốt.

Improving the performance **–** cải thiện*:* sau khi đã đánh giá mô hình, các mô hình đạt độ chính xác không tốt thì cần được train lại, mô hình sẽ lặp lại từ bước 3, cho đến khi đạt độ chính xác như kỳ vọng. Tổng thời gian của 3 bước cuối rơi vào khoảng 30% tổng thời gian thực hiện.

1. Các thuật toán máy học cơ bản

* Neural networks

Mạng thần kinh, còn được gọi là mạng thần kinh nhân tạo (ANN) hoặc mạng thần kinh mô phỏng (SNN), là một tập hợp con của học máy.

Neural networks là một chuỗi các thuật toán cố gắng nhận ra các mối quan hệ cơ bản trong một tập hợp dữ liệu thông qua một quá trình bắt chước cách thức hoạt động của bộ não con người. Theo nghĩa này, mạng lưới thần kinh đề cập đến các hệ thống tế bào thần kinh, có bản chất hữu cơ hoặc nhân tạo.

Neural networks là thuật toán được dùng trong việc nhận dạng các mẫu và đóng một vai trò quan trọng trong các ứng dụng bao gồm dịch ngôn ngữ tự nhiên, nhận dạng hình ảnh, nhận dạng giọng nói và tạo hình ảnh.

* Linear regression

Hồi quy tuyến tính là một loại phân tích thống kê được sử dụng để dự đoán mối quan hệ giữa hai biến. Nó giả định mối quan hệ tuyến tính giữa biến độc lập và biến phụ thuộc và nhằm mục đích tìm ra đường phù hợp nhất mô tả mối quan hệ. Đường này được xác định bằng cách giảm thiểu tổng bình phương chênh lệch giữa giá trị dự đoán và giá trị thực tế.

Hồi quy tuyến tính thường được sử dụng trong nhiều lĩnh vực, bao gồm kinh tế, tài chính và khoa học xã hội, để phân tích và dự đoán xu hướng trong dữ liệu. Nó cũng có thể được mở rộng thành hồi quy tuyến tính bội, trong đó có nhiều biến độc lập và hồi quy logistic, được sử dụng cho các vấn đề phân loại nhị phân.

* Logistic regression

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Ví dụ: giả sử bạn muốn đoán xem khách truy cập trang web của bạn sẽ nhấp vào nút thanh toán trong giỏ hàng của họ hay không. Phân tích hồi quy logistic xem xét hành vi của khách truy cập trước đây, chẳng hạn như thời gian dành cho trang web và số lượng các mặt hàng trong giỏ hàng. Quá trình phân tích này xác định rằng, trước đây, nếu khách truy cập dành hơn năm phút trên trang web và thêm hơn ba mặt hàng vào giỏ hàng, họ sẽ nhấp vào nút thanh toán. Nhờ vào thông tin này, sau đó, hàm hồi quy logistic có thể dự đoán hành vi của một khách mới truy cập trang web.

* Clustering

Là nhiệm vụ chia dân số hoặc điểm dữ liệu thành một số nhóm sao cho các điểm dữ liệu trong cùng một nhóm giống với các điểm dữ liệu khác trong cùng một nhóm và không giống với các điểm dữ liệu trong các nhóm khác. Về cơ bản, nó là một tập hợp các đối tượng trên cơ sở sự giống nhau và không giống nhau giữa chúng.

* Decision trees

Là thuật toán được sử dụng để dự đoán giá trị số (hồi quy) và phân loại dữ liệu. Decision trees sử dụng một chuỗi phân nhánh của các quyết định được liên kết có thể được biểu diễn bằng sơ đồ cây. Một trong những ưu điểm của decision trees là chúng dễ xác thực và kiểm tra, không giống thuật toán Neural networks.

* Random forests

Trong một khu rừng ngẫu nhiên, thuật toán máy học dự đoán một giá trị hoặc danh mục bằng cách kết hợp các kết quả từ một số cây quyết định.

* Support Vector Machines (SVM)
* Top of Form

Support Vector Machines (SVM) là một thuật toán máy học phổ biến được áp dụng cho cả bài toán phân loại và hồi quy. Mục tiêu của SVM là tìm ra ranh giới phân chia tối ưu giữa các nhóm dữ liệu.

Trong bài toán phân loại, SVM cố gắng tạo ra một ranh giới phân chia (gọi là siêu phẳng) sao cho khoảng cách từ các điểm dữ liệu gần nhất tới ranh giới này (còn gọi là margin) là lớn nhất. Các điểm dữ liệu gần nhất này được gọi là "vector hỗ trợ" (support vectors), từ đó xuất phát tên gọi của thuật toán.

SVM hoạt động hiệu quả trong các không gian chiều cao (high-dimensional space), có khả năng xử lý các tập dữ liệu có số lượng biến lớn. Nó cũng thích hợp cho các tập dữ liệu lớn, nhưng cần cân nhắc kỹ lưỡng khi lựa chọn các tham số quan trọng như loại kernel (ví dụ: linear, polynomial, radial basis function) để phù hợp với dữ liệu cụ thể và tránh overfitting.

* Naive Bayes Classifier

Naive Bayes Classifier, hay còn được gọi là Phân loại Bayes Đơn giản, là một mô hình phân loại trong máy học, được xây dựng dựa trên định lý Bayes với giả định "naive" (ngây thơ). Điều này ngụ ý rằng các biến độc lập với nhau, mặc dù trong thực tế thường không phải như vậy. Tuy nhiên, mặc dù có giả định này, mô hình Naive Bayes vẫn có hiệu suất tốt trong nhiều tình huống thực tế.

Cách hoạt động của Naive Bayes Classifier dựa trên tính toán xác suất để đánh giá xác suất của một điểm dữ liệu thuộc vào từng lớp phân loại. Dựa trên các xác suất này, điểm dữ liệu được phân loại vào lớp có xác suất cao nhất.

Mô hình Naive Bayes thường được áp dụng trong các bài toán như phân loại văn bản, phân loại email spam, phát hiện cảm xúc trong văn bản, và nhiều ứng dụng khác trong xử lý ngôn ngữ tự nhiên và phân loại dữ liệu. Mặc dù đơn giản, nhưng Naive Bayes có thể hoạt động hiệu quả và đem lại kết quả tốt cho nhiều loại dữ liệu.

* K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) là một thuật toán máy học không giám sát được sử dụng cho các bài toán phân loại và hồi quy. Trong K-NN, việc phân loại hoặc dự đoán giá trị mới được thực hiện dựa trên việc xem xét các điểm dữ liệu gần nhất trong tập huấn luyện.

Cụ thể, khi phải phân loại một điểm dữ liệu mới, K-NN tính toán khoảng cách giữa điểm dữ liệu này với tất cả các điểm trong tập huấn luyện. Sau đó, nó chọn ra K điểm gần nhất và dùng đa số phiếu (ví dụ: phân loại dựa trên lớp phổ biến nhất trong K điểm) để quyết định lớp của điểm dữ liệu mới.

Thuật toán K-NN không yêu cầu việc huấn luyện mô hình trước, mà chỉ cần lưu trữ tất cả các điểm dữ liệu trong tập huấn luyện để sau đó sử dụng cho việc phân loại hoặc dự đoán. Tuy nhiên, việc tính toán khoảng cách với mỗi điểm dữ liệu trong tập huấn luyện có thể làm tăng độ phức tạp tính toán của thuật toán khi số lượng điểm dữ liệu lớn.

1. **Ứng dụng của Machine learning**

Một số ứng dụng của **learning machine** như sau:

Speech recognition (Nhận diện giọng nói)*:* Nhận dạng giọng nói cho phép máy tính , ứng dụng và phần mềm hiểu và dịch dữ liệu giọng nói của con người thành văn bản cho các giải pháp kinh doanh . Mô hình nhận dạng giọng nói hoạt động bằng cách sử dụng trí tuệ nhân tạo (AI) để phân tích giọng nói và ngôn ngữ của bạn , xác định bằng cách học các từ bạn đang nói, sau đó xuất những từ đó với độ chính xác phiên âm dưới dạng nội dung mô hình hoặc dữ liệu văn bản trên màn hình.

Customer service (Chăm sóc khách hàng)*:* Chatbots trực tuyến đang thay thế các tác nhân con người trong hành trình của khách hàng, thay đổi cách nghĩ về sự tương tác của khách hàng trên website và nền tảng xã hội.

Computer vision (Thị giác máy tính): Công nghệ AI này cho phép máy tính lấy thông tin có ý nghĩa từ video, hình ảnh kỹ thuật số và các đầu vào trực quan khác, sau đó thực thi hành động thích hợp.

Recommendation engines (Công cụ gợi ý): Sử dụng dữ liệu hành vi tiêu dùng trong quá khứ, các thuật toán**AI learning** có thể giúp khám phá các xu hướng dữ liệu có thể được sử dụng để phát triển các chiến lược cross-sell hiệu quả hơn.

Fraud detection (Phát hiện gian lận)*:* Các ngân hàng và các tổ chức tài chính có thể sử dụng máy học để phát hiện các giao dịch đáng ngờ.

Machine learning được ứng dụng cực kỳ nhiều trong đời sống hiện nay trong mọi lĩnh vực:

* Tài chính – ngân hàng
* Sinh học
* Nông nghiệp
* Tìm kiếm, trích xuất thông tin
* Tự động hóa
* Robotics
* Hóa học
* Mạng máy tính
* Khoa học vũ trụ
* Quảng cáo
* Xử lý ngôn ngữ tự nhiên
* Thị giác máy tính

Và còn rất rất nhiều lĩnh vực mà machine learning có thể được áp dụng, machine learning tỏ ra cực kỳ hiệu quả, hơn hẳn con người trong cụ thể các lĩnh vực mà chúng được áp dụng.

Một ví dụ đơn giản về sự ứng dụng của Machine Learning là trong lĩnh vực dự báo thời tiết. Trong quá trình dự báo thời tiết, các chuyên gia sẽ sử dụng các phép tính và quan sát từ các thông tin về thời tiết trong quá khứ để dự đoán các điều kiện thời tiết trong tương lai. Tuy nhiên, khi có một lượng lớn quan sát đến hàng triệu hoặc hàng tỉ, con người sẽ gặp khó khăn trong việc xử lý và tính toán trên dữ liệu lớn đó. Hơn nữa, việc xử lý dữ liệu lớn như vậy có thể dẫn đến các sai sót trong dự đoán.

Ứng dụng Machine Learning trong việc dự báo thời tiết giúp máy tính học từ các quan sát được thu thập trong quá khứ. Nhờ vào khả năng xử lý lượng lớn dữ liệu một cách nhanh chóng và hiệu quả, máy tính có thể tạo ra dự đoán về thời tiết trong tương lai với độ chính xác cao hơn rất nhiều so với con người.

Điều này làm nổi bật sự quan trọng và hiệu quả của Machine Learning trong thời đại công nghệ 4.0. Việc hiểu và áp dụng Machine Learning không chỉ mang lại lợi thế lớn cho cá nhân mà còn đóng góp tích cực vào sự phát triển toàn cầu trong lĩnh vực công nghệ.

1. **XỬ LÝ NGÔN NGỮ TỰ NHIÊN**
2. Định nghĩa về xử lý ngôn ngữ tự nhiên

Xử lý ngôn ngữ tự nhiên là việc sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) để phân tích và thao tác ngôn ngữ của con người trong bối cảnh tính toán. Điều này có thể bao gồm các nhiệm vụ như dịch ngôn ngữ, nhận dạng giọng nói, tạo văn bản, phân tích tình cảm và hơn thế nữa. NLP là một lĩnh vực đang phát triển nhanh chóng, với một loạt các ứng dụng trong các lĩnh vực như trí tuệ nhân tạo, truy xuất thông tin và tương tác giữa con người và máy tính.

Các tổ chức ngày nay có khối lượng lớn dữ liệu thoại và văn bản từ nhiều kênh liên lạc khác nhau như email, tin nhắn văn bản, bảng tin trên mạng xã hội, tệp video, tệp âm thanh và nhiều hơn nữa. Họ sử dụng phần mềm NLP để tự động xử lý dữ liệu này, phân tích ý định hoặc cảm xúc trong tin nhắn và phản hồi bằng người thật theo thời gian thực.

Xử lý ngôn ngữ tự nhiên (NLP) đóng một vai trò quan trọng trong việc hiệu quả hóa phân tích dữ liệu văn bản và giọng nói. Công nghệ này có khả năng xử lý sự đa dạng trong phương ngữ, tiếng lóng và những biến thường về ngôn ngữ thường gặp trong các cuộc trò chuyện hàng ngày. Các công ty tích hợp NLP vào các quá trình tự động, bao gồm:

* Xử lý, phân tích và lưu trữ lượng lớn tài liệu văn bản.
* Phân tích phản hồi từ khách hàng hoặc ghi âm cuộc gọi tới tổng đài.
* Triển khai chatbot cho dịch vụ khách hàng tự động.
* Trả lời các câu hỏi về người, sự vật, thời gian và địa điểm.
* Phân loại và trích xuất thông tin từ văn bản.

Ngoài ra, còn có thể tích hợp NLP trực tiếp vào các ứng dụng tương tác với khách hàng để tăng cường khả năng giao tiếp hiệu quả. Ví dụ, một Chatbot có thể phân tích và phân loại các truy vấn từ khách hàng, tự động trả lời các câu hỏi thường gặp và chuyển hướng các truy vấn phức tạp đến bộ phận hỗ trợ khách hàng. Sự tự động hóa này không chỉ giảm chi phí mà còn giúp nhân viên tránh mất thời gian vào các truy vấn lặp lại, từ đó cải thiện mức độ hài lòng của khách hàng.

1. Các thành phần trong xử lý ngôn tự nhiên

Về cơ bản, có hai thành phần chính bên trong một hệ thống NLP, gồm:

* **Hiểu ngôn ngữ tự nhiên** (Natural Language Understanding - NLU)

Hiểu ngôn ngữ tự nhiên là một phần quan trọng của NLP, tập trung vào việc phân tích ý nghĩa ẩn sau các câu. NLU cho phép phần mềm tìm kiếm các câu đồng nghĩa hoặc xử lý các từ có nhiều nghĩa khác nhau.

Trong phần này, các chủ đề cơ bản về NLU như ánh xạ từ đầu vào sang biểu diễn có ý nghĩa và phân tích các khía cạnh khác nhau của ngôn ngữ.

* **Tạo ngôn ngữ tự nhiên** (Natural Language Generation - NLG)

Tạo ngôn ngữ tự nhiên tập trung vào việc tạo ra văn bản hội thoại giống như con người dựa trên từ khóa hoặc chủ đề cụ thể. Quá trình này bao gồm:

* Lập kế hoạch văn bản, trong đó tìm kiếm và truy xuất nội dung từ cơ sở tri thức.
* Lập kế hoạch về câu bằng cách chọn câu từ cần thiết để thiết lập giọng điệu.
* Hiện thực hóa văn bản, ánh xạ kế hoạch thành cấu trúc câu có ý nghĩa.

1. Cách hoạt động của xử lý ngôn ngữ tự nhiên

* Tokenization

Tokenization là quá trình việc rút gọn từ và khôi phục nguyên thể từ bằng cách đơn giản hóa các từ về dạng gốc của chúng. Chẳng hạn, quá trình này có thể biến từ "starting" thành "start".

Loại bỏ từ dừng đảm bảo rằng các từ không đóng góp ý nghĩa đáng kể cho câu, như "cho" và "với", sẽ được loại bỏ khỏi văn bản.

* Morphological analysis

Phân tích hình thái là quá trình chia nhỏ một câu hoặc đoạn văn thành các đơn vị ngôn ngữ nhỏ nhất có ý nghĩa và xác định hình thái, loại từ và ý nghĩa của từng đơn vị đó.

* Parsing

Trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), sử dụng các quy tắc ngữ pháp để xác định xem văn bản đầu vào có hợp lệ về mặt cú pháp hay không, đồng thời giúp hiểu nghĩa của văn bản đó. Quá trình này không chỉ phân tích cú pháp của văn bản mà còn kiểm tra lỗi cú pháp, do đó còn được biết đến là phân tích cú pháp.

* Named Entity Recognition (NER)

Nhận dạng thực thể được đặt tên (NER) có thể được coi là một quá trình làm cho máy có thể nhận dạng các đối tượng theo lớp của chúng và các thông số kỹ thuật khác. Ngoài ra, với thông tin về nhận dạng thực thể được đặt tên này, chúng tôi đã thảo luận về cách chúng tôi có thể triển khai NER bằng cách sử dụng thư viện spaCy và NLTK để nhận dạng con người, địa điểm, tổ chức, quốc gia, thời gian và sự kiện.

* Sentiment analysis

Phân tích cảm xúc (hoặc khai thác ý kiến) là một kỹ thuật xử lý ngôn ngữ tự nhiên (NLP) được sử dụng để xác định xem dữ liệu là tích cực, tiêu cực hay trung tính. Phân tích tình cảm thường được thực hiện trên dữ liệu văn bản để giúp doanh nghiệp theo dõi tình cảm thương hiệu và sản phẩm trong phản hồi của khách hàng , đồng thời hiểu được nhu cầu của khách hàng.

* Machine Translation

Machine translation là quá trình dịch tự động văn bản từ ngôn ngữ tự nhiên này sang ngôn ngữ tự nhiên khác bằng ứng dụng máy tính. Điều này có nghĩa là bạn thêm văn bản vào phần mềm dịch máy bằng ngôn ngữ nguồn và để công cụ tự động chuyển văn bản sang ngôn ngữ đích đã chọn.

* Question Answering

Question Answering (QA) là một lĩnh vực trong xử lý ngôn ngữ tự nhiên (NLP) tập trung vào phát triển hệ thống hoặc mô hình có khả năng hiểu và trả lời câu hỏi từ người dùng một cách tự động. Mục tiêu của QA là giúp máy tính hiểu câu hỏi được đặt và trả lời chúng bằng cách trích xuất thông tin từ nguồn dữ liệu có sẵn.

* Information Retrieval-based QA: Sử dụng kỹ thuật truy xuất thông tin để tìm kiếm và trích xuất câu trả lời từ nguồn dữ liệu lớn.
* Machine Learning-based QA: Sử dụng mô hình học máy để dự đoán câu trả lời dựa trên dữ liệu đào tạo.

1. Ứng dụng của xử lý ngôn ngữ tự nhiên

Nhận dạng chữ viết: Có hai loại nhận dạng, ví dụ như loại đầu tiên là nhận dạng bản in: nhận dạng và chuyển đổi văn bản sách giáo khoa. Khó hơn nữa là nhận dạng chữ viết tay, vốn khó vì chữ viết tay không có mẫu rõ ràng và thay đổi tùy theo từng người. Với chương trình nhận dạng xuất bản, hàng ngàn cuốn sách thư viện có thể được chuyển đổi thành tài liệu điện tử trong thời gian ngắn. Nhận dạng chữ viết tay của con người có ứng dụng trong tội phạm học và bảo mật thông tin (nhận dạng chữ ký điện tử).

Nhận dạng tiếng nói: Nhận biết âm thanh và chuyển đổi chúng thành văn bản tương ứng. Giúp mọi người sử dụng thiết bị nhanh hơn và dễ dàng hơn. Đây cũng là bước đầu tiên cần thực hiện để hiện thực hóa giao tiếp giữa con người và robot. Một ví dụ về ứng dụng nhận dạng giọng nói đang giúp đỡ người khiếm thị.

Tổng hợp tiếng nói: Từ văn bản được tự động tổng hợp thành giọng nói. Thay vì phải tự đọc nội dung của một cuốn sách hoặc trang web, nó sẽ tự động đọc. Giống như nhận dạng giọng nói, tổng hợp giọng nói là trợ thủ đắc lực cho người mù nhưng ngược lại, nó là bước cuối cùng trong quá trình giao tiếp giữa robot và con người.

Dịch tự động (Machine translate)*:* là một chương trình dịch tự động từ ngôn ngữ này sang ngôn ngữ khác. Nghiên cứu và ứng dụng dịch máy hiện nay. Phát triển nhanh chóng, ứng dụng có độ chính xác cao.

Tìm kiếm thông tin (Information retrieval): Hãy hỏi và chương trình sẽ tự động tìm nội dung phù hợp nhất. Thông tin đang bùng nổ, đặc biệt là với sự phát triển của Internet khiến việc lấy thông tin trở nên dễ dàng hơn bao giờ hết. Vấn đề là tìm được thông tin phù hợp và đặc biệt thông tin đó phải đáng tin cậy. Các công cụ tìm kiếm trên Internet như Google hay Yahoo hiện chỉ phân tích nội dung rất đơn giản dựa trên tần suất từ khóa, thứ hạng trang và nhiều tiêu chí đánh giá khác để đưa ra kết luận. Kết quả là nhiều lượt tìm kiếm không nhận được câu trả lời phù hợp, thậm chí dẫn đến các liên kết không liên quan do các website dùng thủ đoạn lừa đảo để giới thiệu sản phẩm (kỹ thuật SEO - Search Engine Optimization). Trên thực tế, cho đến nay chưa có công cụ tìm kiếm nào hiểu được ngôn ngữ tự nhiên của con người, ngoại trừ www.ask.com, được coi là “hiểu biết”; trả lời các câu hỏi có cấu trúc đơn giản nhất.

Tóm tắt văn bản: Từ một văn bản dài được tóm tắt thành một văn bản ngắn hơn nhưng vẫn chứa những nội dung thiết yếu nhất.

Khai phá dữ liệu và phát hiện tri thức*:* là tìm kiếm thông tin mới trong nhiều tài liệu khác nhau. Khai thác dữ liệu mô phỏng việc học tập của con người và tìm thấy thông tin hữu ích. Ở mức độ đơn giản với các công cụ tìm kiếm, nó cho phép bạn đặt câu hỏi theo cách mà công cụ tìm kiếm tìm ra câu trả lời phù hợp nhất.

Ví dụ ứng dụng của NLP:

Social Listening:

Trong thời đại cạnh tranh ngày càng gia tăng, việc lắng nghe khách hàng trên các kênh xã hội để hiểu xu hướng và thực hiện các chiến dịch tiếp thị phù hợp là rất cần thiết. Tuy nhiên, việc phân tích hàng nghìn, hàng nghìn bình luận của khách hàng trên các fan page, website… chưa bao giờ là điều dễ dàng. Chưa kể kết quả phân tích kém chính xác vì có quá nhiều thư rác, dữ liệu rác.

Những hạn chế của hệ thống Social Listening cũ sẽ không còn là nỗi lo với công nghệ xử lý ngôn ngữ tự nhiên thuần Việt của EM&AI:

Lọc dữ liệu rác/spam: tối ưu chất lượng dữ liệu, tăng hiệu suất xử lý và độ chính xác của kết quả phân tích.

Phân tích Intent/Entity/Sentiment: phát hiện và thông báo những biểu hiện bất thường, vấn đề, scandal mà doanh nghiệp có thể gặp phải.

Trích xuất Insight, báo cáo tự động: thông qua Entity Classification, hệ thống nhận diện những chủ đề được đề cập thường xuyên và đang hình thành xu hướng của cộng đồng mạng.

Helpdesk:

Bộ phận trợ giúp là bộ phận hỗ trợ dịch vụ của một công ty có vai trò cung cấp thông tin và hỗ trợ giải quyết các thắc mắc của khách hàng. Helpdesk hiện đang được sử dụng ở nhiều công ty thuộc bộ phận chăm sóc khách hàng. Phần mềm hỗ trợ khách hàng giúp các công ty giải quyết các yêu cầu của khách hàng hiệu quả hơn bằng hệ thống quản lý yêu cầu. Tuy nhiên, nếu số lượng vé quá lớn sẽ gây ra các vấn đề sau:

* Nhân viên không thể xử lý được tất các yêu cầu của khách hàng
* Không nhận diện được yêu cầu được ưu tiên xử lý.
* Nhân viên xử lý không đúng chuyên môn.
* Mất nhiều thời gian xử lý những vấn đề lặp lại.

Công nghệ xử lý ngôn ngữ tự nhiên thuần Việt có thể giúp ích được cho hệ thống Helpdesk của doanh nghiệp với những tính năng như:

* Tự động phân tích yêu cầu với công nghệ xử lý ngôn ngữ tự nhiên tiếng Việt sau đó phân loại, ưu tiên yêu cầu dựa trên phân tích cảm xúc (Sentiment Analytics).
* Tự động định tuyến yêu cầu của khách hàng đến nhân viên có đúng chuyên môn.
* Hỗ trợ phản hồi nhanh yêu cầu của khách hàng.
* Đừng để khách hàng quay lưng với doanh nghiệp bạn để tìm đến những dịch vụ tốt hơn. Hãy chuẩn bị những công cụ để mang lại cho khách hàng một dịch vụ chất lượng nhất.

Document Management System:

Nhiều công ty đau đầu trong việc tìm kiếm thông tin giữa hàng chục nghìn tài liệu được in và lưu trữ. Ngoài ra, việc không quản lý hoạt động in ấn của công ty dẫn đến lãng phí giấy in.

Doanh nghiệp có thể tiết kiệm tới 80% thời gian khôi phục tài liệu/văn bản bằng cách duy trì kiểm soát chặt chẽ hoạt động in ấn của công ty thông qua hệ thống quản lý tài liệu.

Một số tính năng của hệ thống Document Management có thể kể đến như:

* Phân tích nội dung: ứng dụng NLP và AI trong phân tích nội dung văn bản.
* Tóm lượt nội dung: trích xuất những nội dung quan trọng giúp quản lí và hỗ trợ truy xuất nhanh chóng.
* Bổ sung nội dung: cho phép truy cập vào văn bản trực tuyến, bổ sung thông tin, mã code mà không cần file gốc.
* Phân loại nội dung: phân loại văn bản theo nội dung chính, giúp sắp xếp và trích xuất dễ dàng nhanh chóng.
* Mã hóa chủ đề: chuyển đổi chủ đề dưới dạng QR code/barcode phục vụ cho việc quản lí, trích xuất dữ liệu/văn bản in ấn.

Hệ thống Document Management có thể kết nối với máy in vật lí hiện tại của doanh nghiệp để sử dụng những tính năng trên. Ứng dụng cao cho hầu hết doanh nghiệp thường xuyên in ấn và lưu trữ với khối lượng lớn.

Có thể kể tới một số bài toán xử lý ngôn ngữ tự nhiên trong phát triển chatbot bao gồm:

* Bài toán xác định ý định người dùng (intent detection)
* Bài toán trích xuất thông tin (named entity extraction)
* Quản lý hội thoại (dialogue management)
* Mô hình sinh hội thoại cho chatbot.

1. **PHÂN LOẠI VĂN BẢN**
2. Định nghĩa về văn bản

Khái niệm văn bản là một khái niệm cơ bản trong lĩnh vực ngôn ngữ học và trí tuệ nhân tạo, đề cập đến một đoạn văn bản hoặc một tập hợp các từ, cụm từ và câu được sắp xếp theo một trật tự nhất định để truyền đạt ý nghĩa hoặc thông điệp. Dưới dạng dữ liệu, văn bản thường được biểu diễn bằng các chuỗi ký tự hoặc các token, trong đó mỗi token có thể là một từ, một cụm từ, một ký tự hoặc một số ký tự đặc biệt.

Chi tiết hơn, một văn bản có thể bao gồm các thành phần sau:

Từ và Cụm Từ: Là các đơn vị cơ bản của văn bản, bao gồm các từ đơn (như "nhà", "đi", "là") hoặc các cụm từ (như "nhà hàng", "điều kiện thời tiết", "trong khi").

Câu: Là một chuỗi các từ và cụm từ được sắp xếp theo một cách cú pháp nhất định để truyền đạt một ý nghĩa hoặc một thông điệp. Mỗi câu thường kết thúc bằng dấu chấm câu, dấu phẩy, hoặc các dấu câu khác.

Đoạn: Là một nhóm các câu liên quan nhau, thường chứa một ý chính hoặc một phần của thông điệp được truyền đạt trong văn bản.

Tài liệu Văn Bản: Là một tập hợp các đoạn văn bản và các phần khác nhau của văn bản, có thể là một bài báo, một trang web, một email, một cuốn sách, v.v.

Ngôn Ngữ và Cú Pháp: Văn bản được viết bằng một ngôn ngữ cụ thể và tuân thủ các quy tắc cú pháp của ngôn ngữ đó, bao gồm cú pháp ngữ pháp, cú pháp từ vựng, và cú pháp ngữ cảnh.

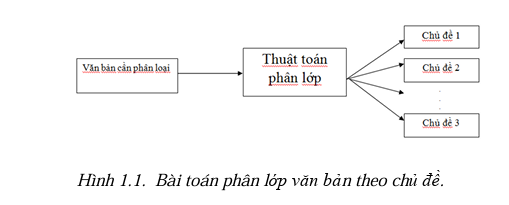
Văn bản không chỉ là một nguồn thông tin quan trọng mà còn là đối tượng nghiên cứu và xử lý quan trọng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), máy học, khai phá dữ liệu, và nhiều lĩnh vực khác.

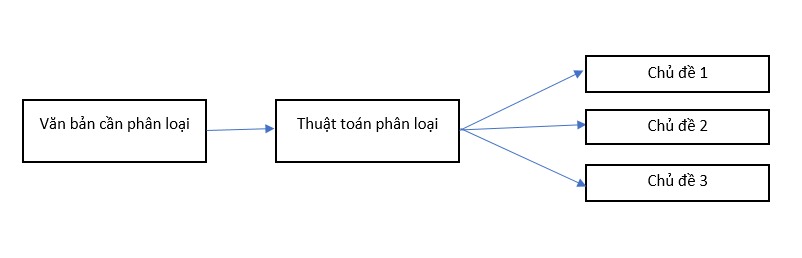
1. Định nghĩa về phân lớp văn bản

Phân loại văn bản là quá trình tự động gán nhãn (tên lớp / nhãn lớp) cho các văn bản ngôn ngữ tự nhiên vào một hoặc nhiều lớp được xác định trước. Đây là một bước quan trọng trong khai phá dữ liệu văn bản, nơi các văn bản được phân loại dựa trên nội dung của chúng vào các chủ đề đã được xác định trước.

Phân loại văn bản có nhiều ứng dụng, bao gồm hỗ trợ trong quá trình tìm kiếm thông tin, chiết lọc thông tin, lọc văn bản hoặc tự động dẫn đường cho các văn bản tới các chủ đề xác định trước. Quá trình này có thể được thực hiện thủ công hoặc tự động bằng cách sử dụng các kỹ thuật học máy có giám sát.

Các hệ thống phân loại văn bản có thể được áp dụng trong nhiều ngữ cảnh, như phân loại tài liệu trong thư viện điện tử, phân loại các bài báo trên các trang tin điện tử, và nhiều ứng dụng khác. Khi được triển khai hiệu quả, các hệ thống này có thể cung cấp kết quả đáng kể, giúp tối ưu hóa công việc và tạo ra lợi ích đáng kể cho người sử dụng.



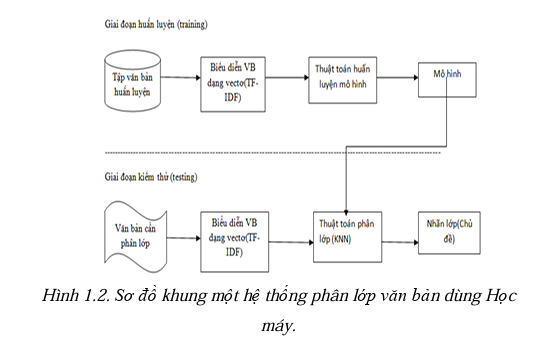


1. Mô hình phân loại văn bản

Phân loại văn bản là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), trong đó mục tiêu là tự động gán các nhãn hoặc danh mục cho các văn bản dựa trên nội dung của chúng. Đây là một bước quan trọng trong khai phá dữ liệu văn bản, giúp tổ chức và hiểu được các tài liệu văn bản một cách tự động và hiệu quả. Dưới đây là các bước cơ bản của bài toán phân loại văn bản:

* Thu thập dữ liệu: Bước đầu tiên là thu thập các tài liệu văn bản cần phân loại từ các nguồn khác nhau như trang web, cơ sở dữ liệu, tập tin văn bản, v.v.
* Tiền xử lý dữ liệu: Dữ liệu văn bản thường cần được tiền xử lý để loại bỏ các thành phần không cần thiết như dấu câu, stop words (các từ phổ biến như "và", "là", "ở"), và thực hiện việc chuẩn hóa văn bản như việc chuyển đổi chữ in thường thành chữ in hoa.
* Biểu diễn văn bản: Các văn bản sau khi đã được tiền xử lý cần được biểu diễn dưới dạng các vector số hóa để có thể được sử dụng trong các mô hình máy học. Các phương pháp phổ biến bao gồm TF-IDF (Term Frequency-Inverse Document Frequency) và Word Embeddings (như Word2Vec, GloVe).
* Xây dựng mô hình: Tiếp theo, một mô hình máy học hoặc học sâu (deep learning) được huấn luyện bằng dữ liệu huấn luyện đã được biểu diễn để dự đoán nhãn cho các văn bản mới. Các mô hình phổ biến bao gồm Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, và Neural Networks.
* Đánh giá mô hình: Mô hình được đánh giá bằng cách sử dụng các phương pháp đánh giá như cross-validation hoặc holdout validation để đảm bảo hiệu suất của nó trên dữ liệu kiểm tra.
* Tinh chỉnh và triển khai: Nếu cần thiết, mô hình được tinh chỉnh và cải thiện trước khi triển khai vào sản phẩm hoặc ứng dụng thực tế.

Bài toán phân loại văn bản có ứng dụng rộng rãi trong nhiều lĩnh vực như phân loại email (spam vs. non-spam), phân loại tin tức, phát hiện cảm xúc trong văn bản, và tổ chức thông tin trong các hệ thống quản lý tri thức.

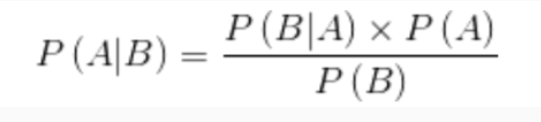


1. **Naive Bayes**
2. Định nghĩa về Naive Bayes

Naive Bayes Classification (NBC) là một phương pháp dựa trên Định lý Bayes về xác suất, được sử dụng để đưa ra dự đoán và phân loại dữ liệu dựa trên thông tin quan sát và thống kê từ dữ liệu. NBC là một trong những thuật toán phổ biến trong lĩnh vực Machine Learning, được áp dụng rộng rãi để tạo ra các dự đoán chính xác từ tập dữ liệu thu thập được. Thuật toán này được đánh giá cao vì tính dễ hiểu và độ chính xác cao của nó. NBC thuộc nhóm thuật toán Học có giám sát (Supervised Machine Learning Algorithms), nghĩa là nó học từ các ví dụ được cung cấp từ các mẫu dữ liệu đã biết trước.

ĐỊNH LÝ:

Định lý Bayes cho phép tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra. Xác suất này được ký hiệu là P(A|B), và đọc là “xác suất của A nếu có B”. Đại lượng này được gọi xác suất có điều kiện hay xác suất hậu nghiệm vì nó được rút ra từ giá trị được cho của B hoặc phụ thuộc vào giá trị đó



Theo định lí Bayes, P(A|B) sẽ phụ thuộc vào 3 yếu tố:  
- Xác suất xảy ra A của riêng nó, không quan tâm đến B. Kí hiệu là P(A).  
- Xác suất xảy ra B của riêng nó, không quan tâm đến A. Kí hiệu là P(B).  
- Xác suất xảy ra B khi biết A xảy ra. Kí hiệu là P(B|A). Đại lượng này gọi là khả năng (likelihood) xảy ra B khi biết A đã xảy ra.

1. Cách hoạt động của Naive Bayes

Cách hoạt động của thuật toán Naive Bayes dựa trên định lý Bayes về xác suất. Thuật toán này giả định rằng các đặc trưng đầu vào độc lập với nhau, tức là sự xuất hiện của một đặc trưng không phụ thuộc vào sự xuất hiện của các đặc trưng khác. Dựa trên giả định này, Naive Bayes tính toán xác suất của một lớp dựa trên các đặc trưng của mẫu dữ liệu.

Cách hoạt động như sau:

* + Xác định các đặc trưng: Đầu tiên, thuật toán cần xác định các đặc trưng hoặc thuộc tính của dữ liệu. Ví dụ, trong bài toán phân loại email, các đặc trưng có thể bao gồm từ vựng xuất hiện trong email, số lượng từ trong email, hoặc các đặc điểm của email như độ dài, tỷ lệ các ký tự in hoa, v.v.
  + Tính toán xác suất của mỗi lớp: Naive Bayes tính toán xác suất của mỗi lớp dựa trên các đặc trưng của mẫu dữ liệu. Đối với mỗi lớp, nó tính toán xác suất có điều kiện của mỗi đặc trưng dựa trên dữ liệu huấn luyện.
  + Áp dụng định lý Bayes: Sau khi tính toán xác suất có điều kiện cho mỗi lớp, Naive Bayes áp dụng định lý Bayes để tính toán xác suất của mỗi lớp dựa trên các đặc trưng của mẫu dữ liệu. Điều này giúp dự đoán xác suất của mỗi lớp cho mẫu dữ liệu mới.
  + Chọn lớp có xác suất cao nhất: Naive Bayes chọn lớp có xác suất cao nhất là lớp dự đoán cho mẫu dữ liệu. Nó xác định lớp mà mẫu dữ liệu có xác suất cao nhất thuộc vào và gán lớp đó cho dự đoán cuối cùng.
  + Đưa ra dự đoán: Cuối cùng, thuật toán đưa ra dự đoán bằng cách gán lớp có xác suất cao nhất cho mẫu dữ liệu. Điều này có nghĩa là nó xác định lớp mà mẫu dữ liệu được phân loại vào và đưa ra dự đoán tương ứng.

1. Ưu điểm và nhược điểm của Naive Bayes

Ưu điểm:

* Dễ thực hiện và nhanh chóng: Thuật toán Naive Bayes thực hiện phân loại một cách dễ dàng và nhanh chóng. Vì vậy, nó thích hợp cho các bài toán có thời gian đáp ứng nhanh hoặc cần sự đơn giản.
* Hội tụ nhanh hơn so với các mô hình khác: Naive Bayes thường hội tụ nhanh hơn các mô hình phân biệt khác như hồi quy logistic, giúp giảm thời gian huấn luyện.
* Yêu cầu ít dữ liệu đào tạo: Thuật toán Naive Bayes yêu cầu ít dữ liệu đào tạo hơn so với một số mô hình khác, điều này có thể rất hữu ích khi dữ liệu huấn luyện có hạn.
* Khả năng mở rộng cao: Naive Bayes có khả năng mở rộng tốt và chia tỷ lệ tuyến tính với số lượng đặc trưng dự đoán và điểm dữ liệu.
* Dự đoán xác suất và xử lý dữ liệu liên tục và rời rạc: Thuật toán Naive Bayes có thể đưa ra dự đoán xác suất, giúp người dùng hiểu rõ hơn về độ tin cậy của kết quả phân loại. Ngoài ra, nó cũng có khả năng xử lý cả dữ liệu liên tục và rời rạc.
* Phân loại cả hai bài toán nhị phân và đa lớp: Naive Bayes có thể được sử dụng cho cả hai loại bài toán phân loại: nhị phân (hai lớp) và đa lớp (nhiều lớp). Điều này làm cho nó trở thành một lựa chọn linh hoạt cho nhiều loại bài toán.

Nhược điểm:

* Giả định độc lập: Một trong những giả định chính của Naive Bayes là giả định về độc lập giữa các đặc trưng khi biết lớp. Trong thực tế, các đặc trưng thường không độc lập hoàn toàn, điều này có thể dẫn đến các dự đoán không chính xác nếu có sự tương quan giữa các đặc trưng.
* Tần số không: Nếu một đặc trưng trong dữ liệu test không xuất hiện trong tập dữ liệu huấn luyện, Naive Bayes sẽ gán xác suất của nó bằng 0, dẫn đến việc mô hình không thể tạo ra dự đoán cho những trường hợp này.
* Dự đoán không chính xác khi dữ liệu phức tạp: Naive Bayes thường hoạt động tốt trên dữ liệu đơn giản và có cấu trúc rõ ràng. Tuy nhiên, khi dữ liệu phức tạp và có tương quan cao giữa các đặc trưng, Naive Bayes có thể tạo ra các dự đoán không chính xác.
* Dễ bị ảnh hưởng bởi dữ liệu nhiễu: Naive Bayes có thể bị ảnh hưởng bởi dữ liệu nhiễu hoặc các đặc trưng không quan trọng. Điều này có thể dẫn đến việc mô hình không hoạt động tốt trên dữ liệu thực tế.
* Không phản ứng tốt với biến đầu vào liên tục: Naive Bayes thường không phản ứng tốt với các biến đầu vào liên tục và yêu cầu việc chuyển đổi các biến này thành các biến rời rạc trước khi huấn luyện mô hình.

1. Ứng dụng của Naive Bayes

* Phân loại văn bản: Naive Bayes thường được sử dụng trong phân loại văn bản, chẳng hạn như phân loại email vào thư rác và thư không phải thư rác. Nó cũng có thể được áp dụng trong việc phân loại văn bản theo chủ đề hoặc ngôn ngữ.
* Dự đoán cảm xúc từ văn bản: Thuật toán Naive Bayes cũng được sử dụng để phân tích cảm xúc từ văn bản, như xác định liệu một bình luận trên mạng xã hội là tích cực, tiêu cực hay trung tính.
* Lọc spam: Với khả năng phân loại hiệu quả và tính nhanh nhẹn, Naive Bayes thường được sử dụng trong các hệ thống lọc thư rác để phân biệt giữa email spam và email hợp lệ.
* Hệ thống gợi ý: Naive Bayes có thể được tích hợp vào hệ thống gợi ý để đề xuất sản phẩm hoặc nội dung dựa trên lịch sử hoạt động của người dùng.
* Phân loại dữ liệu y tế: Trong lĩnh vực y học, Naive Bayes có thể được sử dụng để phân loại dữ liệu y tế, như xác định liệu một bệnh nhân có mắc một loại bệnh nhất định hay không dựa trên các dữ liệu y tế.
* Phát hiện gian lận: Naive Bayes cũng có thể được sử dụng trong các hệ thống phát hiện gian lận, chẳng hạn như phát hiện gian lận trong giao dịch tài chính hoặc giao dịch trực tuyến.
* Phân loại tin tức: Thuật toán này có thể được áp dụng trong việc phân loại tin tức theo chủ đề, như phân loại tin tức thể thao, tin tức chính trị, tin tức khoa học, v.v.

1. **KỸ THUẬT TF – IDF**
2. Định nghĩa TF – IDF

**Tần số từ (TF - Term Frequency)**

Term Frequency (TF), hay Tần suất xuất hiện của từ, là số lần một từ xuất hiện trong một văn bản. Do các văn bản có thể có độ dài khác nhau, một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn so với một văn bản ngắn. Do đó, để chuẩn hóa, term frequency thường được chia cho độ dài của văn bản, tức là tổng số từ trong một văn bản.



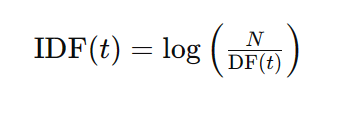
Ví dụ: Nếu từ "machine" xuất hiện 10 lần trong một văn bản có tổng cộng 1000 từ, thì TF của "machine" trong văn bản đó là 10/1000=0,01.

**Tần số văn bản nghịch đảo (IDF - Inverse Document Frequency)**

IDF là viết tắt của "Inverse Document Frequency" (Tần số văn bản nghịch đảo). Đây là một khái niệm quan trọng trong xử lý ngôn ngữ tự nhiên và khai phá dữ liệu văn bản, được sử dụng để đo lường mức độ quan trọng của một từ trong một tập hợp các văn bản.

Cụ thể, IDF đo lường tần suất xuất hiện của một từ trong toàn bộ tập hợp các văn bản và định lượng mức độ quan trọng của từ đó. Mục đích của IDF là tìm ra những từ xuất hiện hiếm trong tập hợp các văn bản, tức là những từ đặc biệt có khả năng đặc trưng và có thể phân biệt các văn bản với nhau.

Công thức tính IDF cho một từ t thường là số lượng văn bản trong tập dữ liệu chia cho Document Frequency (DF) của từ t. Để tránh việc chia cho 0, có thể thêm một hằng số vào mẫu số. Một công thức phổ biến cho IDF là:



Trong đó:

N là tổng số văn bản trong tập dữ liệu.

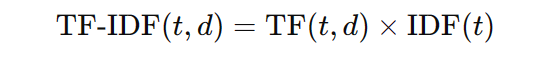
DF(t) là Tần số văn bản của từ t, tức là số lượng văn bản trong tập dữ liệu mà từ t xuất hiện trong đó.

Ví dụ: Nếu trong một tập dữ liệu có tổng cộng 10,000 văn bản, và từ "machine" xuất hiện trong 1,000 văn bản, thì IDF của "machine" là log(10000/1000)= log(10)= 1.

Hàm logarithm được sử dụng để giảm thiểu ảnh hưởng của các từ phổ biến mà không cung cấp nhiều thông tin quan trọng. Càng cao giá trị IDF, tức là càng hiếm khi từ xuất hiện trong các văn bản, càng được coi là quan trọng và có khả năng phân biệt cao giữa các văn bản.

**Tần số từ - Tần số văn bản nghịch đảo (TF-IDF)**

TF-IDF, viết tắt của "Term Frequency – Inverse Document Frequency", là một chỉ số thống kê thể hiện mức độ quan trọng của một từ trong một văn bản so với một tập hợp các văn bản. Giá trị TF-IDF tăng tương ứng với số lần một từ xuất hiện trong văn bản, nhưng thường được điều chỉnh bằng tần số của từ trong toàn bộ tập hợp các văn bản, giúp điều chỉnh thực tế là một số từ xuất hiện thường xuyên hơn nói chung.



Ví dụ: Nếu TF của từ "machine" trong một văn bản là 0.01 và IDF của "machine" trong toàn bộ tập dữ liệu là 1, thì TF-IDF của "machine" trong văn bản đó là 0,01x1= 0,01.

1. Vai trò của TF – IDF

TF-IDF (Term Frequency-Inverse Document Frequency) có vai trò quan trọng trong xử lý ngôn ngữ tự nhiên và khai phá dữ liệu văn bản, và được sử dụng rộng rãi trong nhiều ứng dụng như tìm kiếm thông tin, phân loại văn bản, và trích xuất thông tin. Dưới đây là một số vai trò chính của TF-IDF:

* Xác định từ khóa quan trọng: TF-IDF giúp xác định những từ khóa quan trọng trong một văn bản bằng cách tăng cường trọng số cho những từ xuất hiện ít trong văn bản đó nhưng xuất hiện nhiều trong các văn bản khác.
* Cải thiện hiệu suất tìm kiếm: Trong các hệ thống tìm kiếm, TF-IDF được sử dụng để đánh giá độ phù hợp của một văn bản với một truy vấn tìm kiếm. Các từ có TF-IDF cao hơn được coi là những từ khóa quan trọng và giúp cải thiện chất lượng kết quả tìm kiếm.
* Phân loại văn bản: Trong các bài toán phân loại văn bản, TF-IDF được sử dụng để biểu diễn các văn bản dưới dạng vector có kích thước cố định. Các mô hình học máy sau đó có thể được huấn luyện trên các vector này để phân loại các văn bản vào các lớp khác nhau.
* Trích xuất thông tin: TF-IDF có thể được sử dụng để trích xuất thông tin quan trọng từ các văn bản, như trích xuất từ khóa, phát hiện chủ đề hoặc phân loại văn bản vào các danh mục cụ thể.
* Loại bỏ từ không quan trọng: Các từ có TF-IDF thấp có thể được loại bỏ hoặc được coi là không quan trọng trong các ứng dụng xử lý ngôn ngữ tự nhiên như làm sạch văn bản hoặc giảm kích thước của dữ liệu.

Tóm lại, TF-IDF đóng vai trò quan trọng trong việc biểu diễn và đánh giá mức độ quan trọng của các từ trong văn bản, giúp cải thiện hiệu suất của các hệ thống tìm kiếm và phân loại văn bản, và hỗ trợ trong quá trình trích xuất thông tin từ dữ liệu văn bản.